RSAC 2025 Conference

Many Voices. One Community.

SESSION ID: PDP-W02

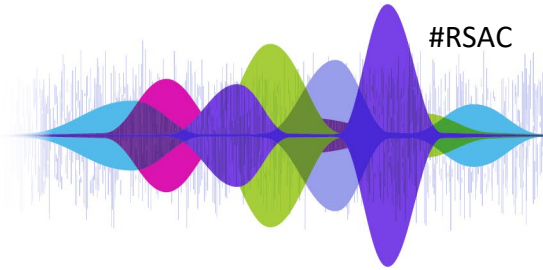
Safe AI Adoption: Standards and Best Practices from Coalition for Secure AI

Ian Brelinsky
Member of Program Staff
OpenAI

Ian's LinkedIn

Josiah Hagen
Security Researcher
Trend Micro

Josiah's LinkedIn



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Agenda

- Why AI Security & Governance Matter
- CoSAI's Role and Approach
- Workstream Highlights
  - AI Supply Chain Security
  - AI Security & Privacy Governance
  - Preparing Defenders for Evolving Threats
  - Secure Design Patterns for Agentic Systems
- Real-World Examples
- Key Findings & Proposed Standards
- Conclusion & Next Steps

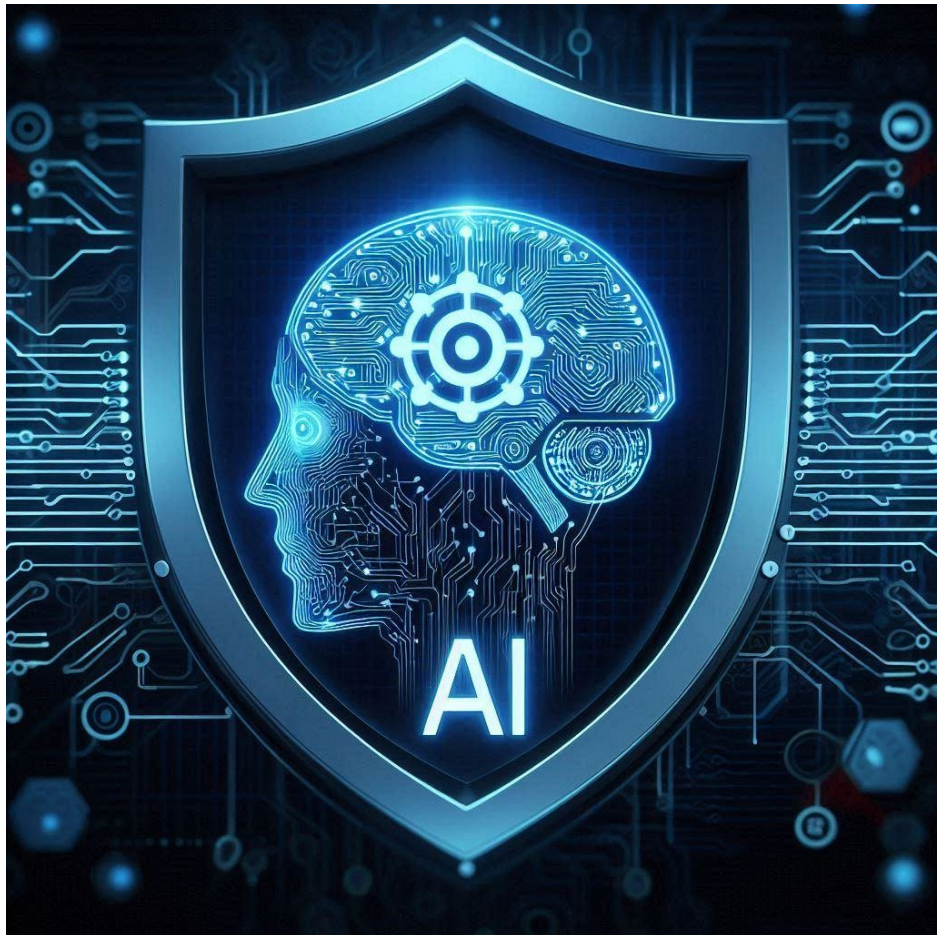
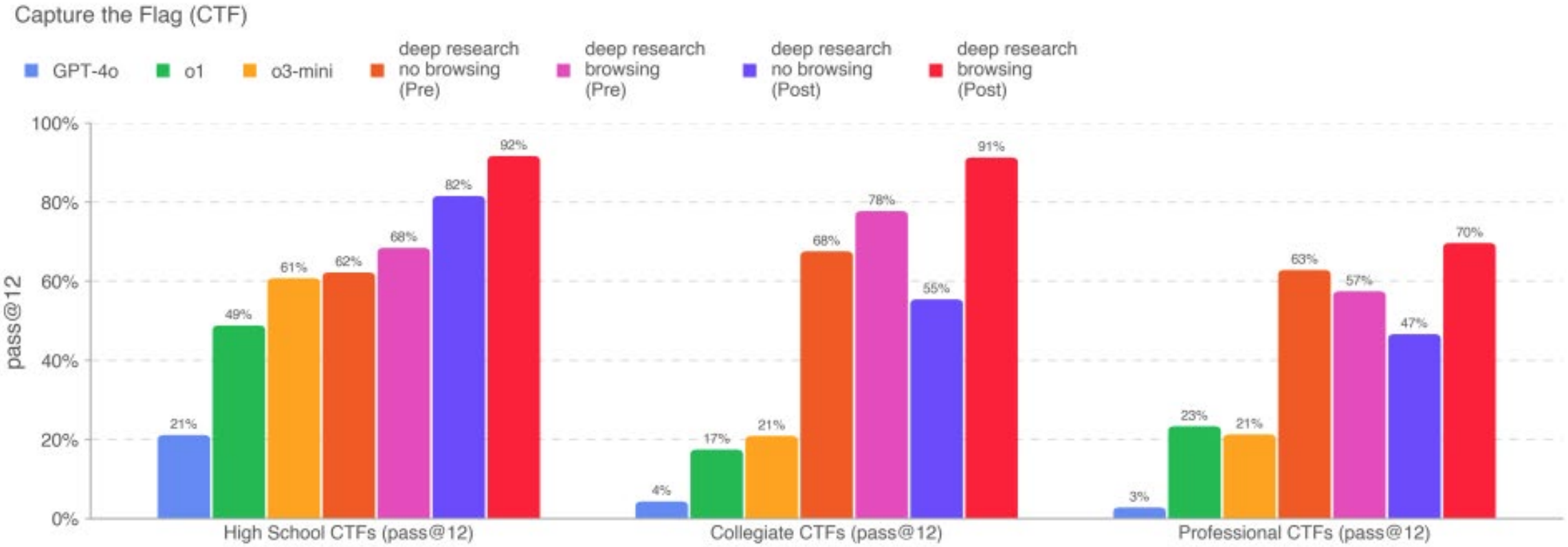


Image Source: OpenAI

Model benchmarks show profound progress



Source: OpenAI

AI is being integrated everywhere

Image Source: OpenAI

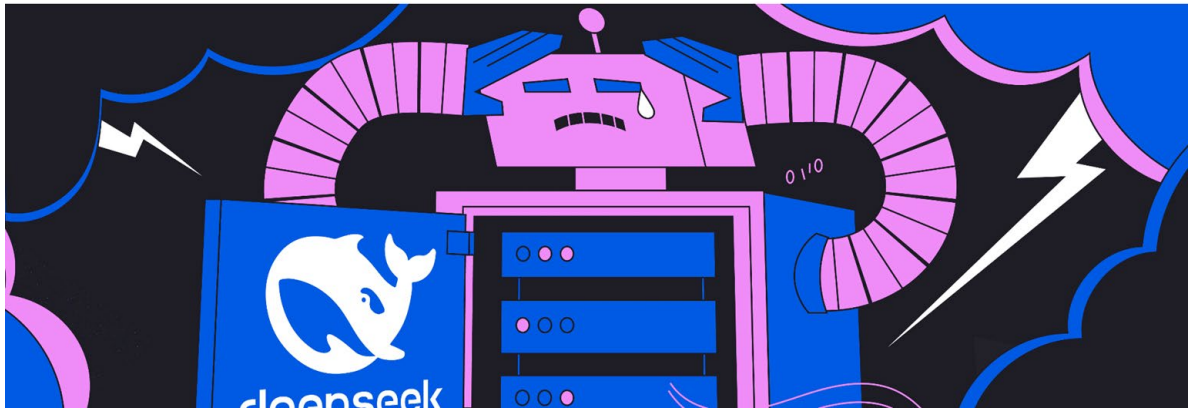
Image: ChatGPT

Security decisions have big impacts

Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History

A publicly accessible database belonging to DeepSeek allowed full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams with highly sensitive information.

Gal Nagli
January 29, 2025
3 minute read



https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak

Secure AI is critical for a secure world

… But, AI can also improve cybersecurity

AI's utility for cybersecurity is improving

Disclosures Classifier

Incident Response

Access Management

Command Line Classification

Code Vulnerability Patching

Much More…

GPT 3.5

Reasoning + Tool Use

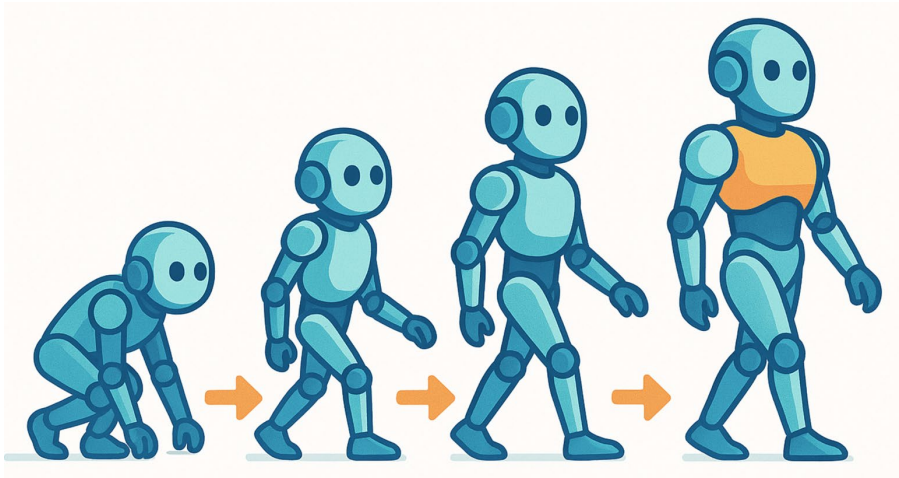


Image Source: OpenAI

Example Risks

Compromptmized (Google): Customized prompt engineering optimizes AI responses.

ConfusedPilot Attack (UT Austin): ConfusedPilot exploits vulnerabilities in RAG-based AI systems.

Influence & Cyber Operations Update (MS): AI is driving influence and cyber operations with emerging risks.

Disrupting Malicious Uses Update (OpenAI): Disrupting malicious AI activities with enhanced safety protocols.

Agentic AI (Trend Micro): Autonomous agentic AI systems present major security risks.

DeepSeek Database Leak (Wiz): DeepSeek app exposed sensitive data and revealed critical security lapses.

Introducing CoSAI

CoSAI unites AI-security experts to share best practices and collaborate.

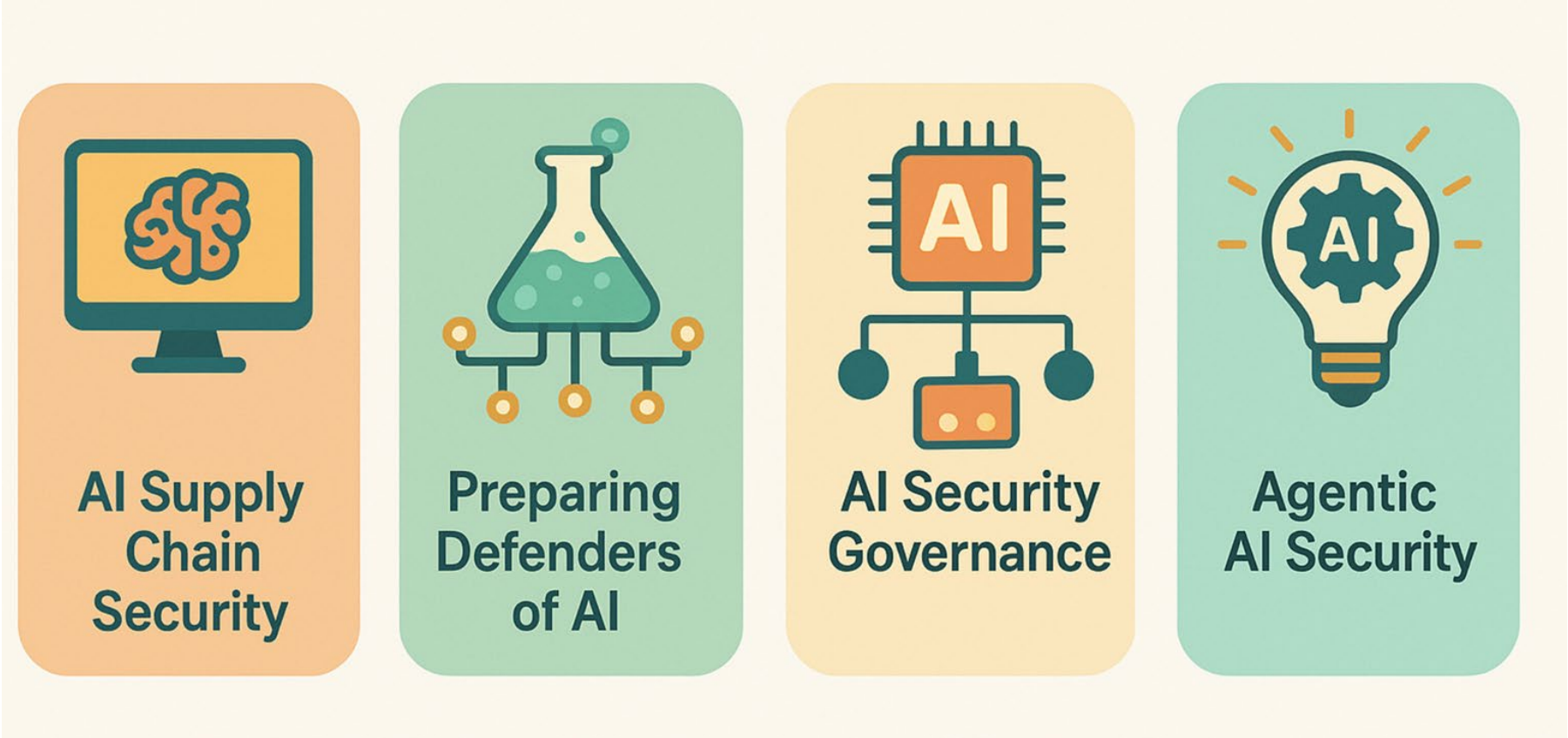


Image Source: OpenAI

AI Supply Chain Security

What's Different & Why It Matters

- **Unique AI Challenges**: AI systems have distinct risks beyond traditional software (e.g., data poisoning, model tampering).
- **Supply Chain Components**: Encompasses Data, Infrastructure, Applications, and Models—each can introduce vulnerabilities.
- **Model Generation & Consumption**: Risks emerge in both stages: from training data integrity to final deployment.
- **Existing Frameworks**: SAIF (Google), MITRE ATLAS, and OWASP AI Exchange highlight AI-specific threats, but gaps remain (e.g., complete supply chain provenance).

AI Supply Chain Security (2)

Key Risks & Mitigations

- **Data Risks**: Poisoned or low-quality data can undermine AI models; practice **data provenance** and access controls.
- **Model Threats**: Model tampering, unsafe serialization formats, and use of unverified dependencies; use **cryptographic signatures** and **secure builds** (SLSA, SBOM).
- **Infrastructure Concerns**: Misconfigurations, hijacked pipelines, overly permissive entitlements—apply **least privilege**, secure notebooks, track model storage.
- **Application & Integration**: Indirect prompt injection, insecure third-party libraries, and poor user input filtering demand **content filtering** and **robust MLOps** practices.

AI Supply Chain Security (3)

Risk Mitigations: Stakeholder Takeaways

- **Executive Leadership**: Enable AI-specific risk governance and cross-functional oversight.
- **Security Practitioners**: Enforce supply chain security controls, monitor dependencies, adopt best practices.
- **AI Engineers/Researchers**: Provide security data in model cards, track model data origins, and collaborate with security teams early.

Preparing Defenders of AI Systems

Evolving AI Security Landscape

- **AI Adoption & Risk**: Rapid enterprise adoption of AI creates new vulnerabilities not covered by traditional security models.

- **Three Core Domains**:

  1. **AI Model Security** – Adversarial threats (prompt injection, data poisoning, etc.)
  2. **AI Platform Security** – Infrastructure, API access, resource hijacking
  3. **AI Security Governance** – Specialized guardrails, audit trails, access control

- **Objective**: Provide practical guidance to help security leaders protect AI assets while fostering responsible innovation.

Preparing Defenders of AI (2)

Gap	What Is Missing
Threat model for AI	Specific adversarial techniques including the mechanisms that exploit weaknesses particular to AI systems
Incident Response or Mitigations	Fine-grain detail on actions for playbooks for incident response, mitigation, and prevention
Integration with other Cybersecurity Frameworks	Map framework details to other frameworks for interoperability
AI Supply Chain Vulnerabilities	Weaknesses in supply chain regarding data, code, and models that create AI system vulnerabilities
Model Behavior Analysis	Determining nominal behavior, monitoring, and analysis of any behavior drift in models to ensure their proper function

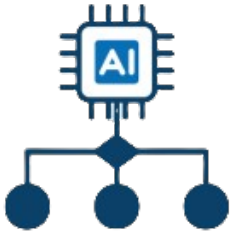
Preparing Defenders of AI Systems (3)

Target Audience	Frameworks
Executives	NIST AI RMF, NIST CSF 2.0, Zero Trust Maturity Model (ZTMM)
CISO/SSO	NIST SP 800-37 (RMF), MITRE ATT&CK, MITRE ATLAS, NIST AI RMF, Zero Trust Maturity Model (ZTMM), OWASP Top 10 for LLMs, STIX
Security Architects	NIST AI RMF, MITRE ATT&CK, MITRE D3FEND, NIST AI AML, MITRE CAPEC, NIST AI RMF, Zero Trust Maturity Model (ZTMM)
IT Architects	NIST AI RMF, Zero Trust Maturity Model (ZTMM), OWASP Top 10 for LLMs, MITRE D3FEND, OCSF
IT Operations	NIST AI RMF, SP 800-37 (RMF), OWASP Top 10 for LLMs, Zero Trust Maturity Model (ZTMM)
SOC Operations	NIST AI AML, MITRE ATT&CK, MITRE ATLAS, STIX, OCSF
Service Operations	NIST AI RMF, SP 800-37 (RMF), OWASP Top 10 for LLMs, Zero Trust Maturity Model (ZTMM)
Auditors/Policy Makers	NIST AI RMF, NIST AI AML
Researchers/Data Scientists	NIST AI AML, MITRE ATT&CK, MITRE ATLAS, MITRE CAPEC, NIST AI RMF, OWASP Top 10 for LLMs
Users/Practitioners	NIST CSF 2.0, OWASP Top 10 for LLMs, NIST AI GAI

Preparing Defenders of AI (4)

Takeaways

- **Build AI Security Foundations**: Extend existing frameworks with AI-specific controls (e.g., prompt filtering, adversarial testing).
- **Adopt a Lifecycle Approach**: Protect AI at every stage—procurement, training, deployment, and monitoring.
- **Roles & Responsibilities**:
  - **Executives & CISOs**: Define governance, strategy, risk appetite
  - **Security Teams**: Implement AI-specific countermeasures, threat modeling, continuous monitoring
  - **Researchers/Data Scientists**: Integrate adversarial robustness, secure data pipelines



AI Security Risk Governance

- Identify and assess security risks of AI systems
- Develop robust security governance frameworks
- Prioritize investments in AI system risk mitigation strategies
- Align with emerging regulatory requirements and best practices

Agentic AI Security

- **Objective:**
  Develop secure design patterns for AI-based agentic systems addressing updated threat models, secure infrastructure, and agent integration.

- **Scope & Focus Areas:**
  - Define CoSAI agentic frameworks & secure AI infrastructure
  - Update threat models for increased agent complexity
  - Create high-level secure design patterns
  - Research secure identity patterns & privileged access

Agentic AI Security (2)

Key Deliverables:

- Agentic-specific security threat model (in collaboration with CoSAI workstreams)
- Conceptual secure design pattern for implementing, protecting, & monitoring agents
- Phase 1: Security-first architectural framework using a single LLM use case
- Secure agentic identity whitepaper detailing design approaches & control invariants

Applying CoSAI's Key Findings

1. Govern your AI Security
   - Determine applicable governance and regulatory policies
2. Secure your AI Supply Chain
   - Assess your supply chain risks and apply the mitigations
3. Prepare your defenders of AI systems
   - Map your role to frameworks and consult the summary

Why CoSAI? Collaborative Advantage

OpenAI
TREND MICRO
EY
Google
IBM
intel
Microsoft
NVIDIA
paloalto NETWORKS
PROTECT AI
PayPal
snyk
zscaler

ADVERSA
aim security
amazon
ANTHROP\C
blinder
chainguard
CipherData
CISCO
cohere
CRANIUM
CYWARE
DELL Technologies
FRONTIERX
GEN LAB
Harvey.
HIDDENLAYER
Invariantlabs
Lasso SECURITY
LEGIT SECURITY
logitech
Lenovo
magic mirror
mozilla
Opal Security
Operant
Red Hat
STYRK AI
Thomson Reuters
TROJ.AI
VE3
ULYSSEAN
WIZ

Image Source: OpenAI

Filling in the Gaps

- Model Signing
- Model Cards
- Security Levels
- AI CAPEC
- Threat Modeling
- Incident Response for AI
- Zero Trust for AI
- Assessing Risk to AI systems
- Securing Agentic AI

Image Source: OpenAI

Apply this info:

1. Learn the security risks for AI systems
2. Determine enterprise policy for AI services, data and identities
3. Monitor the AI supply chain of data, models and code being used
4. Secure AI systems using enterprise appropriate frameworks
5. Coordinate responsibilities to ensure security through AI lifecycle

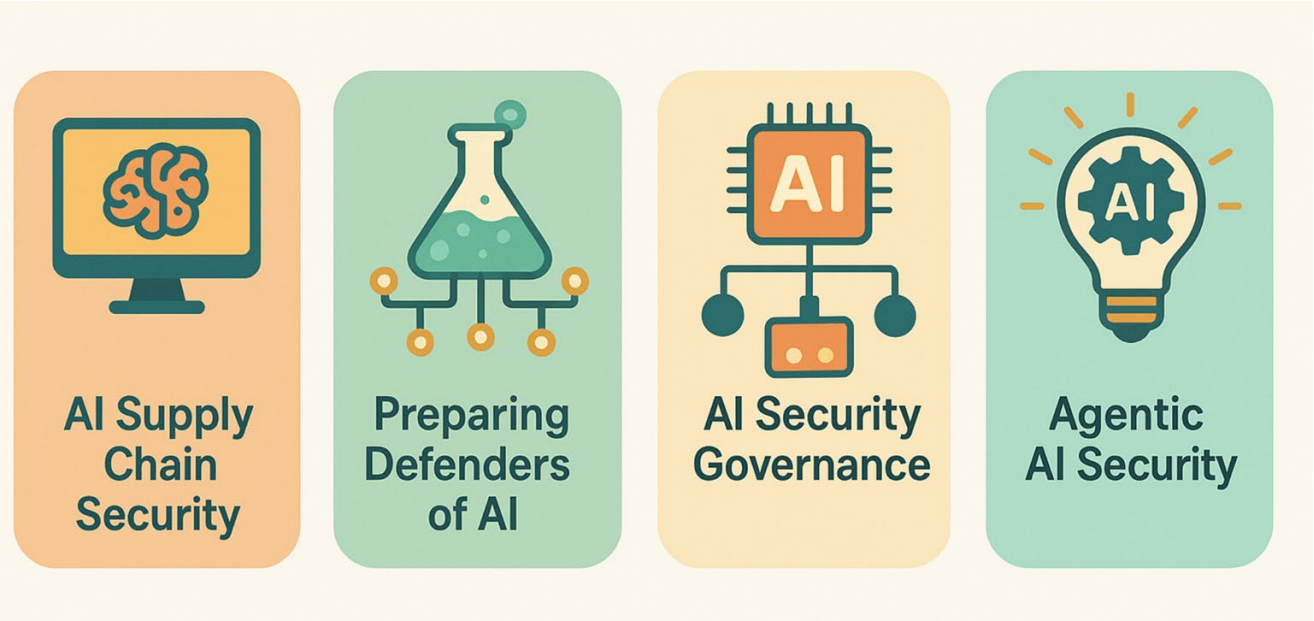


Image Source: OpenAI

Browse our findings & connect with us!

CoSAI is an OASIS Open Project. How you can contribute:

Engage with CoSAI Workstreams

- Monitor & Observe: Join mailing lists & connect via Slack (info on GitHub).
- Actively Contribute: Submit code, documentation, or technical artifacts via GitHub. Contribution details are available here.

Join the Leadership Team

- Help shape governance, strategy, and outreach efforts.
- Secure a seat on the Project Governing Board and Technical Steering Committee.
- Details on how to get involved are [here].

For questions on contributions or sponsorship, contact the OASIS Team.

RSAC 2025 Conference

Many Voices.
One Community.

Thank You

www.coalitionforsecureai.org

https://github.com/cosai-oasis